

0300
#3
LTYSON
02.27.02

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

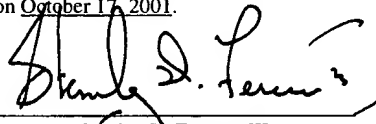
In re Application of : Liqin Shen et al.
Serial No. : 09/944,332 Examiner : Unassigned
Filed : August 30, 2001 Art Unit : Unassigned
For : METHOD AND SYSTEM FOR AUTOMATICALLY
EXTRACTING NEW WORD



October 17, 2001

CLAIM FOR PRIORITY UNDER 35 U.S.C. § 119

I hereby certify that this correspondence and any documents referred to as enclosed therewith are being deposited with the United States Postal Service as first class mail, addressed to the Assistant Commissioner for Patents, Washington, DC 20231 on October 17, 2001.


Stanley D. Ference III
Reg. No. 33,879

October 17, 2001
Date of Signature

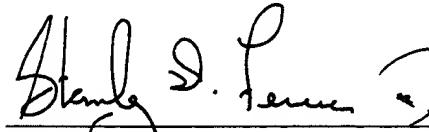
Assistant Commissioner for Patents:
Washington, DC 20231

Sir:

A claim for priority is hereby made under the provisions of 35 U.S.C. § 119 for the above-identified U.S. patent application based upon Chinese patent application

number 00 1 26471.0 filed August 30, 2000. A certified copy of this Chinese patent application is filed herewith.

Respectfully submitted,

A handwritten signature in cursive script, appearing to read "Stanley D. Ference III", written over a horizontal line.

Stanley D. Ference III
Registration No. 33,879

FERENCE & ASSOCIATES
129 Oakhurst Road
Pittsburgh, Pennsylvania 15215
(412) 781-7386
(412) 781-8390 - Facsimile

Attorneys for Applicant

Enclosure



JP9-2000-0191 US
(YOR)

证 明

本证明之附件是向本局提交的下列专利申请副本

申 请 日： 2000 08 30

申 请 号： 00 1 26471.0

申 请 类 别： 发明专利

发明创造名称： 自动新词提取方法和系统

申 请 人： 国际商业机器公司

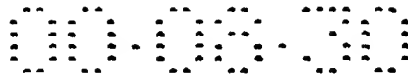
发明人或设计人： 沈丽琴； 施勤； 柴海新



中华人民共和国
国家知识产权局局长

姜 颖

2000 年 12 月 4 日



权 利 要 求 书

- 1、一种自动提取新词的方法，包括步骤：
对原始语料库进行分段，成为分段的语料库；
5 将分段的语料库分割成子串并对子串在语料库中的出现次数进行统计；
过滤掉假词，输出新词。
- 2、如权利要求 1 的方法，其特征在于：
对原始语料库进行分段的步骤包括利用标点符号或阿拉伯
10 数字及字母字符串或新词模板进行分段的步骤。
- 3、如权利要求 1 或 2 的方法，其特征在于：对原始语料库进行分段的步骤还包括利用公共词汇表进行分割的步骤。
- 4、如权利要求 1 或 2 的方法，其特征在于：
对分段的语料库进行分割及统计的步骤包括通过构建 GAST
15 结构进行分割及统计的步骤。
- 5、如权利要求 4 的方法，其特征在于：构建 GAST 结构的步骤还包括限定子串的长度的步骤。
- 6、如权利要求 1，2，4 或 5 的方法，其特征在于：滤除假词的步骤包括：
20 滤除功能词；
滤除那些几乎总是与更长的子串一起出现的子串；以及
滤除其出现次数少于预定阈值的子串。
- 7、如权利要求 1，2，4 或 5 的方法，其特征在于：对原始语料库进行分段的步骤还包括将预先识别出的功能词作为分段符
25 进行处理的步骤。
- 8、如权利要求 3 的方法，其特征在于：对原始语料库进行分段的步骤还包括将预先识别出的功能词作为分段符进行处理的步骤。
- 9、如权利要求 3 的方法，其特征在于：滤除假词的步骤包

括:

滤除功能词;

滤除那些几乎总是与更长的子串一起出现的子串; 以及
滤除其出现次数少于预定阈值的子串。

5

10、一种自动提取新词的系统, 包括:

用于将原始语料库分成分段的语料库的装置;

用于将分段的语料库分割成子串并对子串在语料库中的出
现次数进行统计的装置; 以及

10

用于过滤掉假词, 输出新词的装置。

11、如权利要求 10 的系统, 其特征在于:

用于对原始语料库进行分段的装置包括利用标点符号或阿
拉伯数字及字母字符串或新词模板进行分段的装置。

15

12、如权利要求 10 或 11 的系统, 其特征在于: 对原始语
料库进行分段的装置还包括利用公共词汇表进行分割的装置。

13、如权利要求 10 或 11 的系统, 其特征在于:

对分段的语料库进行分割及统计的装置包括通过构建 GAST
结构进行分割及统计的装置。

20

14、如权利要求 13 的系统, 其特征在于: 构建 GAST 结构
的装置还包括用于限定子串的长度的装置。

15、如权利要求 10, 11, 13, 14 的系统, 其特征在于: 滤
除假词的装置包括:

滤除功能词的装置;

滤除那些几乎总是与更长的子串一起出现的子串的装置;

25

以及

滤除其出现次数少于预定阈值的子串的装置。

16、如权利要求 10, 11, 13 或 14 的系统, 其特征在于:
对原始语料库进行分段的装置还包括将预先识别出的功能词作为
分段符进行处理的装置。

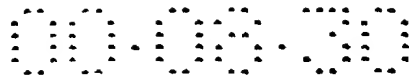
17、如权利要求 12 的系统，其特征在于：对原始语料库进行分段的装置还包括将预先识别出的功能词作为分段符进行处理的装置。

5 18、如权利要求 12 的系统，其特征在于：滤除假词的装置包括：

滤除功能词的装置；

滤除那些几乎总是与更长的子串一起出现的子串装置；以及

滤除其出现次数少于预定阈值的子串装置。



说明书

自动新词提取方法和系统

本发明涉及语言处理技术领域，尤其涉及从语料库中提取新词的方法。

在很多语言处理技术中，词是最基本的。例如，具有不同属性的词汇表是自然语言理解、机器翻译、自动撰写摘要等的基础。为了检索信息，总是用词作为搜索单位来减少检索结果的冗余。在语音识别中，也通常把词作为最低层次的语言信息，并基于词建立语言模型，以解决单字层次上的声觉不确定性。然而，在有些语言例如中文、日文的书面文字中，在词之间不会留有空格，并且对词的构成也没有明确的定义。例如，有些人可能认为“吃东西”是一个词，而另一些人则认为它由两个词“吃”和“东西”组成。一般说来，中文词由一个中文字或多个中文字组成，他们是具有特定意义的基本单位。已经有很多人工收集的词汇表，它们覆盖了不同领域的不同范围。然而要收集这样的词汇表是不容易的。而且，语言在不断地发展，新词也在不断地出现。例如，“互联网”在若干年以前不是一个词，但现在它却作为一个词在广泛地使用。因此，人们迫切需要一种从给定的大量语料中自动提取新词的方法。本发明的目的就是提供一种能够自动从语料库中提取新词的方法。

为了实现以上目的，本发明提供了一种自动提取新词的方法，包括步骤：对原始语料库进行分段，成为分段的语料库；将分段的语料库分割成子串并对子串在语料库中的出现次数进行统计；过滤掉假词，输出新词。

为了实现以上目的，本发明还提供了一种自动提取新词的系统，用于将原始语料库成分段的语料库的装置；用于将分段的语料库分割成子串并对子串在语料库中的出现次数进行统计的装置；以及用于过滤掉假词，输出新词的装置。

图1是本发明的自动新词提取系统的基本原理图。如图1所示，本

发明的系统包括一个分段模块 1, 采用广泛使用的最大匹配方法或统计分段方法或下面将要描述的本发明的分段方法将原始语料库分成单位序列形成分段语料库; 一个 GAST 模块 2, 利用上面的分段语料库构建一个 GAST 并将各子串在原始语料库中出现的次数进行统计; 一个新词提取模块 3, 根据滤波统计和滤波规则滤除子串中的伪词, 从而最后输出新词。各模块的详细操作将在下面分别详细描述。

下面描述如何根据本发明构建通用原子后缀树 (GAST)。

首先定义字符串 $S=u_1, u_2 \dots u_N$, 其中 u_1 是 S 的单位串。定义 $\text{suffix}_i=u_i, u_{i+1}, \dots u_N$ ($1 \leq i \leq N$) 为 S 的后缀串。一个字符串 S 的原子前缀树 (AST) 是带边和叶子的树, 其中每个叶子都与对应 suffix_i 的下标 i ($1 \leq i \leq N$) 相关联。每条边上都标有字符, 从而每条边上只有一个单位串并且这些被标记的边沿从根到下标为 i 的叶子的路径串接起来形成 suffix_i 。图 2 中示出了串 ababc 的 AST 的例子。关于 AST 的构建在由 Lucas Chi Kwong Hui 发表在 1992 年的 Proceedings of the 2nd Symposium on Combinatorial Pattern Matching 第 230 到 243 页的论文 Color Set Size Problem with Application to String Matching 中有详细描述, 这里就不再详述。从 AST 的结构中, 我们可以得到 AST 的每个节点的信息, 包括:

当前节点 (例如: 节点 6)

{
 路径 (将所有被标志的边沿从根到节点 i 的路径串接起来); (节点 6 的路径是 “ab”)
 路径计数 (路径在串中出现的次数); (“ab” 在串 ababc 中出现两次)
 子节点 节点 i , ..., 节点 j ; (节点 8 和节点 9)
 父节点 节点 f ; (节点 3)
 }

串 S (S 的长度= N) 的 AST 可以在一个 $O(N^2)$ 的空间中建立。对那些计数为 n 的节点, 意味着在建立 AST 时一共被使用了 n 次。如果忽

略因节点重复使用而节省下的空间，AST 的大小是： $N(N+1)/2$ 。实际上，这是所有节点的计数和。

AST 的原理可以被扩展到去存储多于一个的输入串。该扩展被称为通用原子后缀树 (GAST)。如果有 M 个长度为 N_i 的串 ($1 \leq i \leq M$)，则

$$\sum_{i=1}^M \frac{N_i(N_i+1)}{2}$$

图 3 示出了包括串 “abca” “bcab” “acbb” 的 GAST 的例子。从 GAST 的树形结构中，我们很容易得到所有子串的列表以及它们在语料库中的出现次数。

下面描述分段边界 (SB) 模板、新词模板和 GAST 所需空间的压缩。

虽然 GAST 是一种能够简洁地表示串的很好的数据结构，将它实际应用于新词提取时还是有一些问题。对于一个很大的语料库，建立相应的 GAST 结构所需空间太大，效率不高甚至于不可行的。

通常我们需要处理几百万到几十亿个字的语料库，从中提取某一新领域中的新词。如果将它们作为一个串输入到 AST，由于需要的空间太大，要构建这样的 AST 是不实际的。

通过定义 SB 模板和新词模板，我们能够将很长的输入串分成较小的部分，从而能够显著地降低空间需求以构建 GAST 和实际实现自动新词提取。

如上所述，对于长度为 N 的串 S 的 AST 的大小为 $\frac{N(N+1)}{2}$ 。如果将串分成 k 个相等的部分，对于具有 k 个输入串的 GAST，其所需的为 $\frac{N}{k} \left(\frac{N}{k} + 1 \right)$ 。节省下来的空间为 $\frac{N^2}{2} \left(1 - \frac{1}{k} \right)$ 。例如，如果一个 10 个符号长的串被分成两个相等的部分，节省下来的 GAST 节点有 25 个。如果一个 20 个字符长的串被分成了 4 个相等的部分，则节省下来的节点有 150 个。

由于目标新词不可能很长，因此正确定义 SB 来将过长串分成

短串而又不丢失很有可能的新词是很关键的。

下面是一些 SB 模板 (SBP) 的定义:

SBP A: 标点符号自然是 SB;

SBP B: 在语料库中的阿拉伯数字和字母是另一类 SB.

5 对于另外的 SBP, 我们考虑两种情况:

1、以基本的公共词汇表为基础, 定义新词模板对子串进行限制。

10 尽管有很多领域并且每个领域都有自己的专门词汇表, 也不管语言的发展有多么迅速, 都有一些基本词汇是在各个领域中都使用着, 例如“因为”, “生活”等。我们可以首先利用公共词汇的词汇表来将语料库分段。分段的语料库将由单字词和多字词组成,

例如,

15 代表着未来生活方式的互联网技术将不再会将弱视和失明者拒之门外。 (1)

其分段结果为

代表着未来生活方式的互联网技术将不再会将弱视和失明者拒之门外。 (2)

20 以 w 表示多字词, 以 c 表示单字词, 上述句子可以表示为,

$w_1 c_1 w_2 w_3 w_4 c_2 c_3 c_4 c_5 w_5 c_6 w_6 c_7 c_8 w_7 c_9 w_8 c_{10} c_{11} c_{12} w_9$

其中, w_3 表示“生活”, c_4 表示“联”, 以此类推。

定义新词模板 (NWP) 如下:

25 NWP A: $c_i c_{i+1} \dots c_j$, 表示所有由单字词组成的串。例如上面句子中的“互联网”。

NWP B: $w_i c_k$ 或者 $c_i w_k$ 或者 $w_i c_k w_{i+1}$ 或者 $c_i w_k c_{i+1}$ 等等, 表示由单字和多字词组合而成的串例如, 上面句子中的“失明者”。

对于模板 $w_i w_{i+1}$, 表示多字词+多字词, 他们通常称为复合词,

一般来讲不会是要找的新词。因此，在多字词之间，我们可以设定 SB。我们称这样的模板为 SBP C。

根据上面的原理分析上面的句子。因为“未来”、“生活”和“方式”都属于已知的多字词，所以“未来”和“生活”的组合是多字词+多字词，同样，“生活”和“方式”的组合也是多字词+多字词，所以可以在“未来”和“生活”及“生活”和“方式”之间设定 SBP C。又因为“生活”是已知的公共词汇，所以可以忽略“生活”这个词并将两个 SBP C 合并。

我们定义“|”来表示 SB，分析后的句子（1）看起来是：
10 代表着未来|方式的互联网技术将不再将弱视和失明者拒之门外|

这表示有两个串：

（1） 代表着未来

（2） 方式的互联网技术将不再将弱视和失明者拒之门外
15 而不是整个句子（1）将会被输入来构建 GAST。

依据同样的准则，可以对这类模板的各种形式根据需要进行进一步细化，以减少 GAST 所需的空间。其中 SBP 和 NWP 的具体定义可以根据不同的需要随时增加或者修改。例如，在其它实施方式中，可以认为只有两个字的多字词加只有两个字的多字词不属于复合词，有可能是新词。根据词的构成分析，本领域的普通技
20 术人员显然可以设计出各种另外的新词模板。这种用 SBP 将原始语料中的句子分割成短串的方法也可以用于其他语言处理的领域。

如果我们使用 30,000 个词作为基本词汇表，当我们分析有
25 3497 个词的信息技术的特定领域词汇时，我们得到了 990 个 NWP A 词和 2507 个 NWP B 词。

利用上面定义的 SBP，我们对信息技术领域的一百万大小的语料库进行了统计，其结果如表 1 所示。

从表 1 可以看出，利用 SBP A, B 和 C，GAST 节点的数目，即

The figure shows three 5x5 dot patterns. The first pattern represents the digit '0', the second represents '1', and the third represents '2'. Each pattern is formed by a specific arrangement of black dots on a white background.

14

5

10

1 的行 5 所示。

与行 2 相比, 节省的空间是 110, 162 个节点.

0. 基本词汇 (词)	SBP	SB 的数目	串的平均长度	GAST 节点的数目
1. 所有中国字	A	297,68	12.46	2,496,219
2. 所有中国字	A+B	38,063	8.22	1,442,366
3. 60K	A+B+C	31,921	4.52	398,220
4. 30K	A+B+C	31,515	4.61	407,522
5. 所有中国字	A+B+Nup=7	38,063	8.22	1,332,204

表 1 IT 领域中 1M 大小的语料库的统计分析

15

利用上面的机制，为自动新词提取而构建 GAST 所需的空间是可以实现/控制的。

20

构建好 GAST 后，就可以如下所述进行新词的提取了。

词的基本定义是那些经常在一起使用的子串。因此，每个节点路径的计数是判定该路径是否是一个新词的基本测量。如果我

们将“新词”定义为一个在语料库中至少出现过 K 次的连续字符串，其中具体的 K 值可以根据选择新词的需要自行设定，例如设定 $K=5$ ，则自动新词提取的基本原理是用上面描述的方法构建一个相应的 GAST，并对其原始路径计数进行修正，然后对于该树内的
5 每一个节点，如果其修正过的节点计数大于等于 K ，则其对应的相应子串是一个所定义的新词。本领域的技术人员将知道如何根据特定的领域，特定的原始语料库的大小等具体因素通过试验或分析来设定合适的阈值。

因为 GAST 的构建方式和特性并不能保证所有获得的新词都是
10 真正合理有用的，所以在本发明的实施方式中还可以采用其它技术来对新词列表进行修剪。这些技术如下所述。

A. 限制功能词

在中文或日文中，有一些词是经常使用的，如“的”，“也”或“了”。这些辅助词通常不能成为一个新词的结尾或者开头部
15 分，不管它们的访问计数有多大。

B. 选取较长的词

在 GAST 中，如果一个节点的计数等于其所有子节点的计数和，同时其所有子节点都已输出，则意味着该节点所对应的相应子串在给定的语料库中几乎从不单独出现，该子串即使其计数大
20 于等于 K 也不认为是一个新词。因为有些词可能单独出现，也可能与别的更长的词一起出现。所以在具体算法中可以每当输出一个较长的词时，将该较长的词所对应的串的子串所对应的所有节点的计数值减去该较长词节点所对应的计数值。若这些子串所对应的节点计数最后还大于阈值，则这些子串除了与较长的子串所
25 对应的词出现外，本身还可能作为一个词出现。

方法 A 和方法 B 可以有效地保证删除的词不是本发明感兴趣的新词。

C. 还可以根据先验概率来建立过滤规则。例如，如果有一个从标准语料库导出的先验统计语言模型，从中我们得到了 P_s

($w_1...w_n$)，它是新提取的词 $NW=w_1...w_n$ 出现的概率，我们可以很容易从当前语料库中计算出 $P_c(w_1...w_n)$ 。如果 $P_c(w_1...w_n)/P_s(w_1...w_n)$ 的值较大，则意味着 NW 在当前语料库中出现的概率比在标准语料库中出现的概率相对较高，它是一个该领域内的真正新词。否则，意味着 NW 的组合在标准领域中已经很普通，所以不是一个新词。

图 4 示出了根据本发明的新词提取方法的一种实现方式。如图 4 所示，流程从方框 401 开始，构建好 GAST，并对 GAST 按宽度优先遍历的节点序列排序 $N_1, N_2, ...N_m$ 。例如，如图 3 所示，排序方式为节点 N_1 为 $1/5$, N_2 为 $2/4$, N_3 为 $3/3$, ... N_{17} 为 $17/1$ 。接着到达方框 402，设定一个控制变量 $s=m$ ，在图 3 的情况下 $m=17$ ，所以 $s=17$ 。接着到达方框 403，看节点 N_s 的计数值是否大于等于阈值 k 。在图 3 的例子中，计数值等于 1，小于阈值（假设阈值大于 1，这是通常的）。所以流程到达方框 410，将控制变量 s 的值减 1，即打算对下一个节点进行处理。接着到达方框 411，判断 s 是否大于 0，即判断是否还有节点待处理。若判断结果为否，则流程到达方框 412 结束。若方框 411 的判断为是，则流程又到达方框 403 进行处理，判断该节点的计数值是否大于阈值。假设这次计数值大于阈值，则流程到达方框 404，判断该节点是否是一个功能词。若判断结果为是，则流程到达方框 410 进行上面已经描述的处理。若方框 404 的判断为否，则到达方框 407，取出该节点对应的路径并作为新词输出。输出新词后，流程到达方框 408，对该新词中的任何一个子串所对应的节点的计数值减去该新词所对应的节点的计数值并写回原处，如方框 409 所示。例如，如果在方框 407 所输出的新词为“日新月异”，则对“日”，“日新”，“日新月”，“新”，“新月”，“新月异”，“月”，“月异”所对应的节点的计数值都减去节点“日新月异”所对应的计数值并写回原处。在方框 405 判断是否所有的子串已经处理完毕，若

所有的子串处理完毕则到达方框 410 接着上面描述的处理。

5 经过以上的处理，我们可以得到一个新词列表。显然上面的流程在具体实现中可以有各种变形。例如，本实施例中是把一个单字也当成可能的新词。在其它的实施例中，如果总是不把单字当成一个新词，则处理流程可以简化。删除单字功能词的步骤也可以不要。

10 本领域的技术人员将会明白，可以对上述的实施方式进行各种改进而不会偏离本发明的范围。例如，如果当前面所述的功能词刚好出现在标点符号的前面或后面时，因为功能词一般不会是词头或词尾，所以可以与标点符号一起当成分段符。利用公共词汇表进行分割可以与限定子串长度进行分割结合使用。

说明书附图

图1

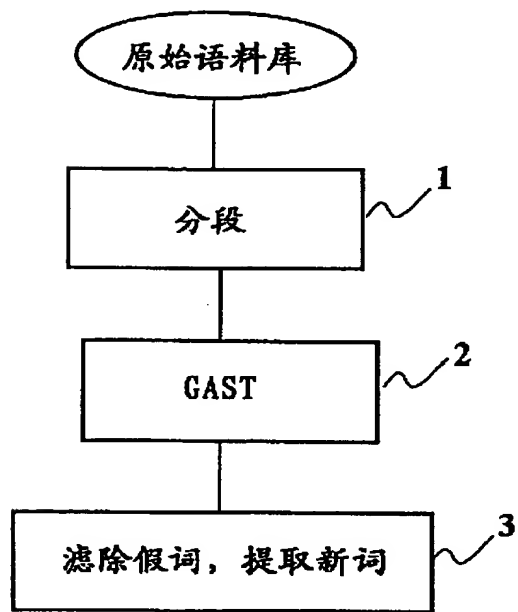


图2

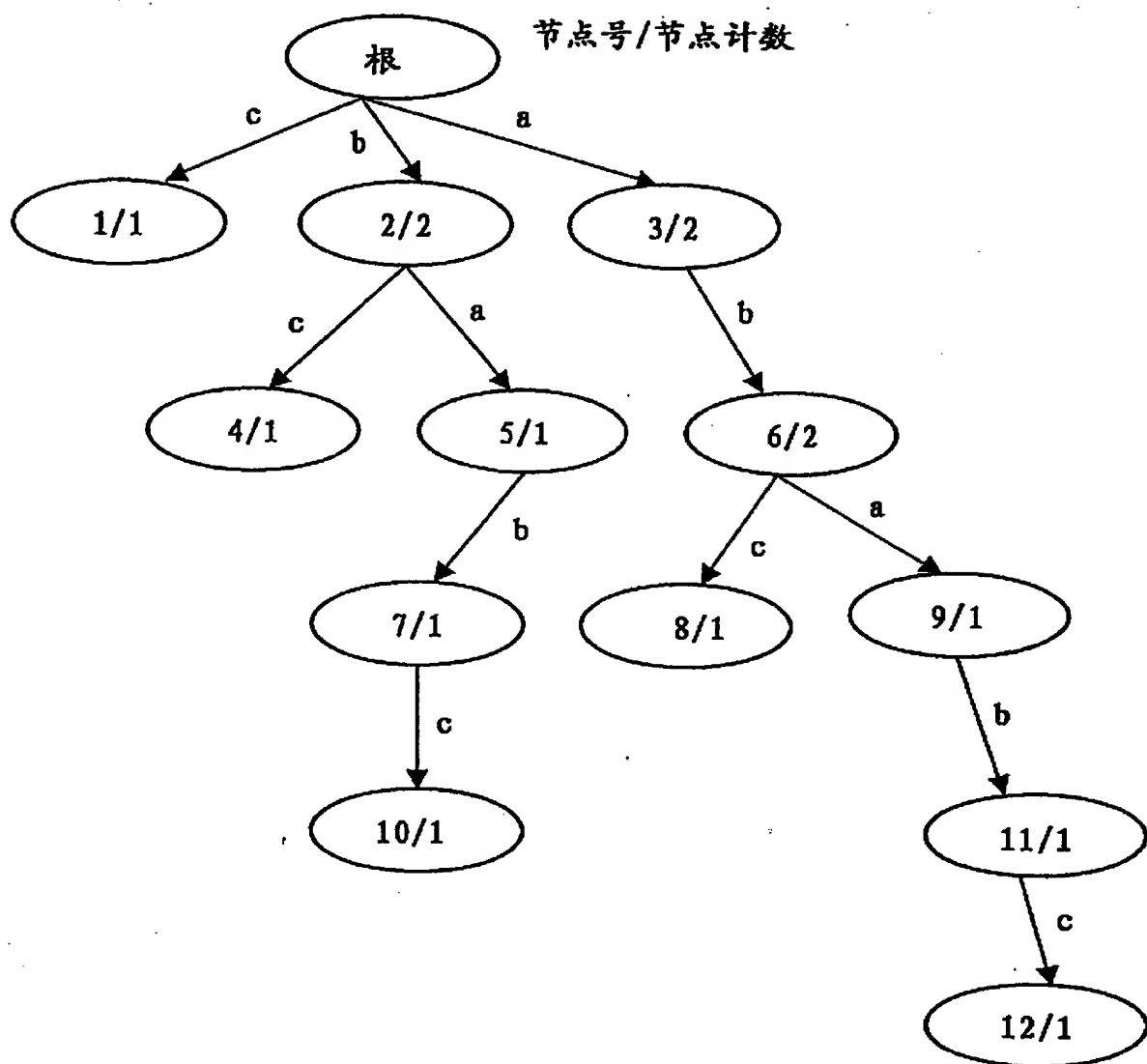


图3

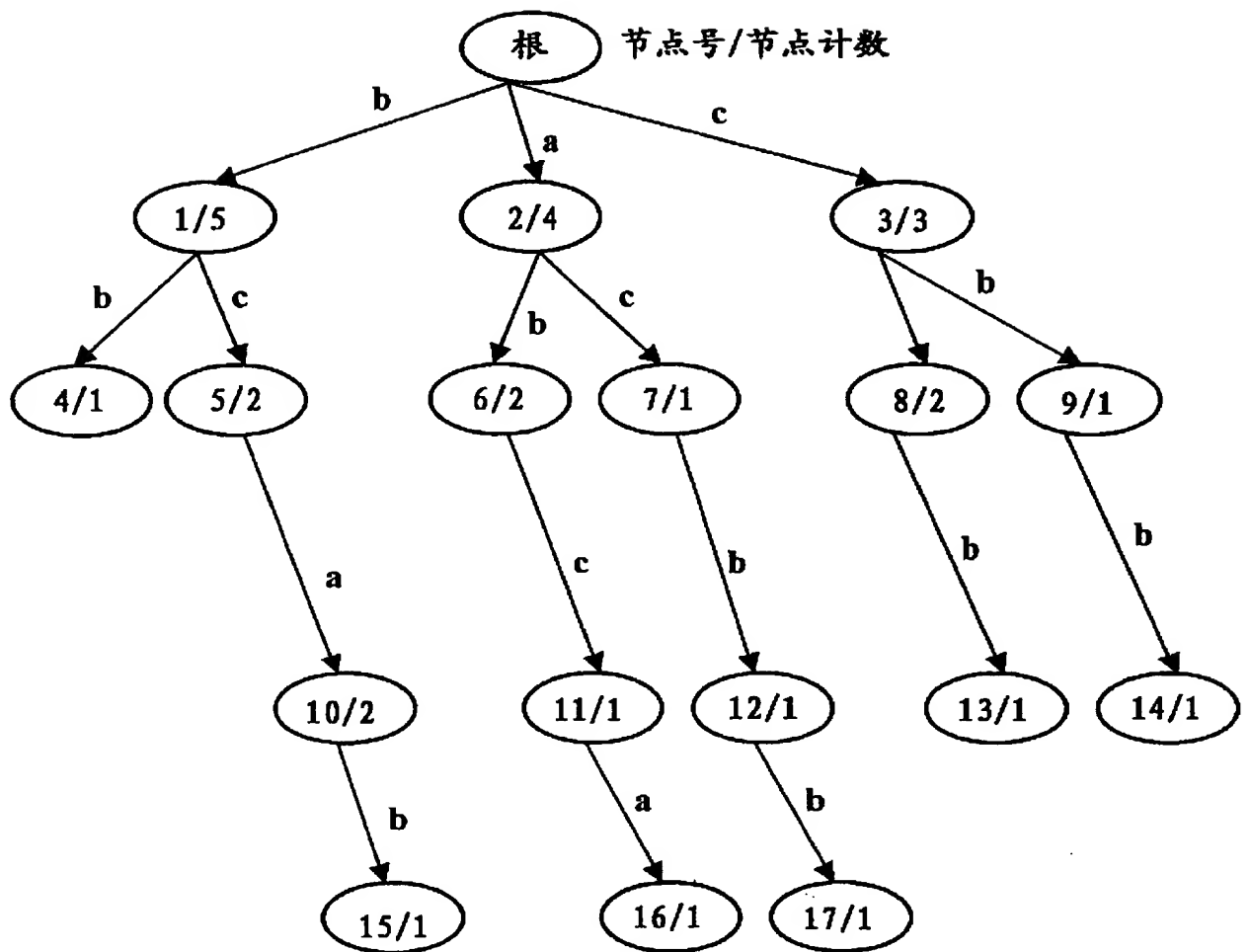


图 4

